CASE STUDY - HEALTHCARE
DATA ANALYSIS

Website: www.analytixlabs.co.in
Email: info@analytixlabsl.co.in

BUSINESS CONTEXT:

The Cloudera Data Science Challenge is a rigorous competition in which candidates must provide a solution to a real-world big data problem that surpasses a benchmark specified by some of the world's elite data scientists.

In the U.S., Medicare reimburses private providers for medical procedures performed for covered individuals. As such, it needs to verify that the type of procedures performed and the cost of those procedures are consistent and reasonable. Finally, it needs to detect possible errors or fraud in claims for reimbursement from providers. You have been hired to analyze a large amount of data from Medicare and try to detect abnormal data -- providers, areas, or patients with unusual procedures and/or claims.

The objective of the Cloudera Data Science Challenge 2 was to uncover anomalous patients, procedures, providers, and regions in the United States government's Medicare health insurance system.

PROBLEM SUMMARY:

The Challenge was divided into the following three parts, each of which had specific requirements that pertained to identifying anomalous entities in different aspects of the Medicare system:

Part 1: Identify providers that overcharge for certain procedures or regions where procedures are too expensive.

- Part 1A: Highest Cost Variation
- Parts 1B: Highest-Cost Claims by Provider
- Parts 1C: Highest-Cost Claims by Region
- Part 1D: Highest Number of Procedures and Largest Differences between Claims and Reimbursements

Part 2: Identify the three providers that are least similar to other providers and the three regions that are least similar to other regions.

- Part 2A: Providers Least Like Others
- Part 2B: Regions Least Like Others

Part 3: Identify 10,000 Medicare patients who are involved in anomalous activities.

DATA AVAILABLE:

Completing the different parts of the Challenge required using several data sources.

Parts 1 and 2 were based on financial summary data from 2011 that were made available by the Centers for Medicare and Medicaid Services (CMS) in both comma-separated values (CSV) formats.

- Medicare_Charge_Inpatient_DRG100_DRG_Summary_by_DRG_FY2011.csv
- Medicare_Charge_Outpatient_APC30_Summary_by_APC_CY2011.csv
- Medicare_Provider_Charge_Inpatient_DRG100_FY2011.csv
- Medicare_Provider_Charge_Outpatient_APC30_CY2011_v2.csv

You can also find these details from the below links.

Inpatient financial summary data:

https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Inpatient.html

Outpatient financial summary data:

https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Outpatient.html

Part 3 of the Challenge required patient demographic information and patient-procedure transaction information

- Patient_history_samp.csv
- Review_patient_history_samp.csv
- Rreview_transaction_coo.csv
- Transaction_coo.csv

Expectations from the trainees:

Trainees should able to import the data into statistical softwares and peform variety of data preprocessing, analysis, and visualization using different techniques.

[Hints:

Part-1 may require performing different data manipulations and descriptive statistics to identify certain procedures or regions where procedures are too expensive. You may also use box-plots and pie-charts to visualize these analyses. Identify claimed Charges for the Three Procedures That Have the Highest Coefficient of Variation (Relative Variation).

Part-2 may require performing clustering, Euclidian distances, linear regression or any other machine learning techniques to identify the three providers that are least similar to other providers and the three regions that are least similar to other regions. You may also use scatter plots to visualize.

Part-3 may require performing association analysis, cluster analysis, graph presentations etc to identify Medicare patients who are involved in anomalous activities]